Predictive Approaches to Discordancy Testing

by

Seymour Geisser
University of Minnesota

Technical Report No. 498
September, 1987

Research supported by National Science Foundation Grant DMS-8601314 and National Institute of Health Grant No. GM 25271.

Predictive Approaches to Discordancy Testing

Seymour Geisser
University of Minnesota

0. Introduction

Simple tests of significance are, by all accounts, the most prevalent of statistical procedures in current use. One need only peruse a journal in which papers using statistical methods are published and one is immediately beset by either varying numbers of stars indicating significance at particular levels or the exact P-values themselves. Why this should be so is rather odd as significance testing has been roundly denounced by most frequentist and Bayesian statisticians alike as generally useless under any circumstances. However there have been a few statisticians who have vigorously defended tests of significance in certain limited situations Fisher (1956), Barnard (1962) and Box (1980).

Clearly, so-called pure significance tests, whether based on frequentist, Bayesian, or predictivist principles, suffer not only from a suitable choice of a critical region but are also criticized on the basis that alternative hypotheses have not been taken into account. From the Bayesian viewpoint (one which we generally subscribe to) this requires knowledge of all possible alternatives and their prior probabilities--a formidable if not impossible task in many instances. There are also situations where reasonable modeling alternatives are very difficult to contemplate. At any rate we believe that there are situations where a simple test of significance can be useful. The area we shall focus on is discordancy testing where a particular situation is such that a standard model has proven useful in similar settings in the past but perhaps in the current experiment one or two observations are discordant from the rest. For example there may be errors in transcribing data, numbers misread, digits transposed or an incorrect sign before a number; or a stipulated experimental condition that did not actually obtain for each of the observables. Therefore the surprise engendered by a small P-value of an appropriate significance test is useful in detecting these and other anomalies. If it is determined that the "discordant" observation(s) make(s) an appreciable difference in a potential inference or decision then this should provide the stimulus for a thorough review of the protocols and procedures in making the observations. A determination then needs to be made as to whether the apparently discordant observable is really incompatible with the rest of the observations or whether the modeling requires revision or both. In simple univariate situations these often take the form of outliers in that one or two observations appear to be distantly removed from the others and discordancy tests for outliers have been devised. For a compendium of frequentist tests the reader is referred to Barnett and Lewis (1978). In this paper we shall use the predictive framework of a Bayesian approach as a basis testing for testing for discordant observations.

We shall present a general framework for such discordancy tests that may simply depend (a) on the identification of a potentially discordant observation because of the intrusion of some untoward event prior to the observance of its actual value or (b) taking into account a diagnostic ransacking of the data in search of potentially discordant observations. Several different approaches are devised and illustrated for certain conventional distributional models. These will involve conditional and unconditional predictive distributions.

1. Testing an Observation Under Suspicious Circumstances

Consider the situation where $Y_1, \ldots, Y_N$ are independently and identically distributed with common distribution function $F_Y(y|\theta)$. Further assume that $g(\theta)$ is a prior probability function for θ. Hence based on this model, say M, one can compute the predictive distribution for a future value Z by calculating

$$F_Z(z|y^{(N)}) = E_\Theta F_Y(z|\theta)$$

where $y^{(N)} = (y_1, \ldots, y_N)$ is the observed set of values of $Y^{(N)} = (Y_1, \ldots, Y_N)$, and the expectation is over the posterior distribution of Θ.

If, when obtaining the experimental value for Y_i, it was noted that some untoward event or suspicious circumstance occurred that may or may not have been related to the experiment, the experimenter could make a determination regarding the inclusion of the experimental outcome by calculating the predictive distribution of Z based on all the observations except for y_i and denoted $y_{(i)}$,

$$F_Z(y_i|y_{(i)}) = \int F_Z(y_i|\theta)dP(\theta|y_{(i)})$$

where $P(\theta|y_{(i)}$ is the posterior distribution of θ given $y_{(i)}$. An assessment of the discordancy of Y_i with $Y_{(i)}$ can be made by calculating a P-value, Geisser (1980)

$$P_i = Pr[Z \varepsilon R|y_{(i)}] \tag{1}$$

for some suitably defined region R using the conditional predictive distribution of Z given $y_{(i)}$. There are several reasonable principles for implementing the above procedure. One would be to order values of the range of Y_i as to their departure from compatibility with the rest of the observations assuming the truth of the original model. Another is to simply calculate

$$P_i = \Pr[f_Z(Z|y_{(i)}) \leq f_Z(y_i|y_{(i)})] \tag{2}$$

In many problems where the predictive distribution is unimodal and symmetric both principles lead to the same test procedure.

This procedure assumes that only y_i is suspect and that the set $y_{(i)}$ concords with the model. This permits one to condition on $y_{(i)}$. If a test is required which does not depend on the assumption that $y_{(i)}$ is concordant then one could calculate the marginal distribution of Y_i,

$$F(y^{(N)}) = \int F(y^{(N)}|\theta)dP(\theta)$$

where $P(\theta)$ is the prior distribution for θ. A computation analgous to (1) can be made using the unconditional distribution of Y_i, namely

$$P_i = \Pr[Y_i \varepsilon R], \tag{3}$$

or

$$P_i = \Pr[f_Z(Z) \le f_Z(y_i)].$$

These calculations have the advantage that they do not depend on an assumption about $Y_{(i)}$, but they can only be used with a proper prior distribution. Of course when the assumption that $Y_{(i)}$ concords can confidently be made, the more sensitive test previously described would be available using the additional information. Further this test exists for certain useful improper prior distributions as well.

Often the statistician may be unaware of the experimental procedures and only when examining the data by plotting or applying analytic diagnostic procedures does the possibility of outliers or discordant observations for the presumed model become evident. Checking with the investigator could reveal some experimental mishap that might have occurred when taking those observations so that appropriate remedies may be undertaken. However, in some cases no such information from the experimenter is available and in these situations it would certainly be wise to take account of the diagnostic checking in computing a significance test. The burden of the rest of this paper is to propose general predictive testing procedures, when the assumption is made that at most there are only a few discordant observations and the data have been ransacked in some predetermined manner. First we shall discuss some ways of ransacking or searching the data for potentially discordant observations from a predictivistic viewpoint.

2. Discordancy Indices.

Before a random sample of independent and identically distributed random variables is taken, all of the potential observables are "created equal" in terms of making inferences or decisions. However, after the observations are in hand this is no longer the case and each can have a different effect on the inference e.g. on the predictive distribution of a future observable (or on the estimation of the set of parameters θ or a subset of them).

Various functions have been proposed to rank the varying influence of observations. For purposes of this paper we shall adhere to those that concern themselves with prediction. Johnson and Geisser (1982, 1983) proposed predictive influence functions (PIF) using the Kullback-Leibler numbers between the predictive distribution of a future value based on $y_{(i)}$ and $y^{(N)} = (y_i, y_{(i)})$. This can also be used in the random sample case to yield an ordering of the discordancy of the observations, i.e. in the absence of other information, the observation yielding the largest Kullback-Leibler number is the foremost candidate for a discordancy test. Thus the PIF, which can be interpreted as a discordancy index here, is

$$I_i = E[\ln f(Z|y_{(i)}) - \ln f(Z|y^{(N)})]$$

where the expectation is taken over the predictive distribution $F_Z(z|y_{(i)})$, i.e. with y_i deleted.

Another such diagnostic or index, Geisser (1980,1985),

$$d_i = f(y_i|y_{(i)})$$

called the Conditional Predictive Ordinate (CPO) ranks the discordancy of observations--the smaller the value of d_i the more discrepant is y_i from $y_{(i)}$, see also Smith and Pettit (1985). Other indices are the unmodified tests (1) and (2) of section 1 which are useful for testing an observable which is initially under suspicion for reasons other than its actual value.

The one given by (2) is somewhat similar to the Aitchison-Dunsmore (1975) atypicality index--the difference being that here each observable Y_i is related to its predictive distribution conditioned on the other observations while their index uses the single predictive distribution which includes y_i.

3. Predictive Discordancy Tests.

In the predictive framework an overall test for model adequacy has been suggested by Box (1980) by calculating for a model M the prior predictive density of the entire set of observables $Y^{(N)} = (Y_1,\ldots,Y_N)$,

$$f(y^{(N)}|M) = \int f(y^{(N)}|\theta,M)g(\theta|M)d\theta,$$

where $g(\theta|M)$ is the prior density of Θ under model M.

He tests the adequacy of M by computing the significance level for the observed value $y^{(N)}$

$$P_y = \Pr[f(Y^{(N)}|M) \le f(y^{(N)}|M)]$$

or for some "checking function" $H(Y^{(N)})$

$$P_h = Pr[f_H(H) \leq f_H(h)]$$

where h is the observed value of H. Small values of P_y or P_h will serve to call model M into question where by model M he means the joint density of $Y^{(N)}$ and Θ i.e.

$$f(y^{(N)},\theta|M) = f(y^{(N)}|\theta,M)g(\theta|M).$$

Sometimes a model may be called into question because of one or more faulty observations. Suppose some checking function, diagnostic criterion, , or discordancy index is used to identify possibly faulty or discordant observations where say H_C chooses y_C as potentially most discordant i.e. the observed value

$$h_C \succ h_i \quad \text{for all } i \neq C$$

where $\succ$ stands for "more discordant." One can check for discordance by calculating a significance level from h_C, the observed value of H_C,

$$Pr[f(H_C|M,C) \leq f(h_C|M,C)]$$

or from

$$Pr[H_C \succ h_C|C,M]$$

where $f(H|M,C)$ is calculated from $f(y^{(N)}|M)$ under choice C i.e. that h_C was the observed value of the most discrepant diagnostic.

To illustrate this, consider a random sample $Y_1,\ldots,Y_N$ from $\theta e^{-\theta y}$ with prior density $g(\theta)$. Suppose the index H happened to lead to choosing as most discordant

$$y_C = \max_i y_i.$$

Then we could calculate

$$\Pr[Y_C \le y_C|M,\theta] = \left(1-e^{-y_C\theta}\right)^N$$
$$\Pr[Y_C \le y_C|M] = \int\left(1-e^{-y_C\theta}\right)^N g(\theta)d\theta$$

and

$$P_C = \Pr[Y_C > y_C|M] = 1 - P(Y_C \le y_C|M).$$

If we have a conjugate prior density for θ with hyperparameters N_0 and $\bar{y}_0$,

$$g(\theta) \propto \theta^{N_0-1} e^{-\theta N_0\bar{y}_0}, \quad \bar{y}_0 > 0, \quad N_0 > 0$$

then the CPO index will still rank y_C the maximal observation as relatively most discordant and

$$P_C = \sum_{j=1}^{N} \binom{N}{j}(-1)^{j+1}\left[\frac{N_0\bar{y}_0}{N_0\bar{y}_0+jy_C}\right]^{N_0} \tag{4}$$

Note that P_C depends only on N, y_C and the hyperparameters of $g(\theta)$, $\bar{y}_0$ and N_0. Incidentally, if we were suspicious of an observation y_i for reasons other that its value as indicated in section (3), an unconditional test, irrespective of $y_{(i)}$, would merely set $N = 1$ and substitute y_i for y_C in (4) yielding

$$P_i = \Pr[y_i \geq y_i|M] = [\frac{N_0\bar{y}_0}{N_0\bar{y}_0+y_i}]^{N_0} .$$

This unconditional test which assumes nothing about the observation set $y_{(i)}$ may be compared to a conditional test derived using (1)

$$P_i = \Pr[Y_i \geq y_i|y_{(i)}, M] = [\frac{N_0\bar{y}_0+N\bar{y}}{N_0\bar{y}_0+N\bar{y}+y_i}]^{N+N_0}$$

which assumes that the set $y_{(i)}$ concords with the model. Another possible unconditional discordancy test alternative to (2) would be to calculate

$$P_C = P[Y_C-Y_{C-1} > y_C-y_{C-1}|M]$$
$$= [\frac{N_0\bar{y}_0}{N_0\bar{y}_0+y_C-y_{C-1}}]^{N_0} , \qquad (5)$$

where y_{C-1} is the second largest observation. This test is completely independent of N but now depends on y_{C-1}. It also appears, as it does in (4), to put relatively heavy stress on the prior.

If the distribution of the index or diagnostic H is independent of θ and the hyperparameters then the prior distribution has no effect whatever on the significance value. This is the usual case for those diagnostics (statistics) that are currently in use and purport to test discordancy for the sampling part of the model rather than the entire model, i.e. including the prior distribution.

Another possible way of looking at this problem is to consider the conditional predictive density

$$f(y_i | y_{(i)}) = \frac{f(y^{(N)}|M)}{f(y_{(i)}|M)}$$

As indicated previously, this is appropriate when suspicion is aroused that something untoward occurred that may have effected the value of a particular observation. A simple predictive significance test as indicated earlier, conditional on $y_{(i)}$ satisfying the model, is

$$P_i = [Pr[Z \varepsilon R_i | y_{(i)}]$$

where R_i is some specified region. If on the other hand the observation is to be examined solely because its observed value was flagged by the diagnostic H as ranked most discordant and no discernible reason for it adduced, than an adjustment to the previous significance test to take account of this fact, at the very least, is in order. One possible way is to reject only at a P that decreases as N increases. How this might be sensibly implemented is not clear. Another possible way of adjusting the test of significance is to calculate

$$P_C = Pr[H > h_C | H > h_{C-1}; y_{(C)}]$$

where h_{C-1} is the value for H which ranks y_{C-1} as the second most discordant observation. For example, assume that the discordancy diagnostic is a monotone increasing function of the observation's value. In this case we suggest, as a simple alternative to finding the significance level from the distribution of the maximum as in (4), using the original predictive distribution of the observable and conditioning it on being greater than the observed value y_{C-1},

$$P_C = \Pr[Z > y_C | Z > y_{C-1}, y_{(C)}].$$

The significance level for the proposed adjusted or Conditional Predictive Discordancy (CPD) test for the exponential problem previously discussed is

$$P_C = \left[\frac{N_0\bar{y}_0 + N\bar{y} + y_{C-1} - y_C}{N_0\bar{y}_0 + N\bar{y}} \right]^{N_0+N-1} \tag{6}$$

which depends not only on the prior hyperparameters N_0 and $\bar{y}_0$ but on the observables y_C, y_{C-1}, $\bar{y}$ and N. We also note that the test exists as N_0-->0 which yields the useful improper prior which purports to reflect little prior information relative to the likelihood.

Discordancy testing procedures can be developed from both the conditional and unconditional approach. There are several aspects of the conditional approach that many statisticians may find appealing. The first one is that the test procedures depend more on the conditional distribution (likelihood) of the observables than the unconditional procedures which depend more heavily on the prior for Θ. The second is that the unconditional approach in general can not be used with an improper prior, while the conditional procedure can be used with certain useful improper priors. A third aspect is that the conditional approach will often lead to much simpler computations than the unconditional approach.

It is to be noted that all these tests are essentially subjective assessments and are not grounded in a frequency theory as there is no attempt to tie this to a class of repetitions. For certain combinations of sampling and prior distributions, some of the CPD tests proposed have conditional and unconditional frequency analogues. For example in (6), $N_0 = 0$ results in the usual improper prior $g(\theta)\alpha\theta^{-1}$ and

$$\lim_{N_0 \dashrightarrow 0} P_i = \left[\frac{N\bar{y} + y_{C-1} - y_C}{N\bar{y}}\right]^{N-1} = P_C \qquad (7)$$

This, it is easy to show, leads to the same significance level one would obtain from the classical frequentists statistic

$$T = \frac{Y_C - Y_{C-1}}{Y_C}$$

used to test for discordance, Dixon (1950, 1951) if one conditioned on

$$U = \frac{Y_C}{N\bar{Y}},$$

so that

$$\Pr[T > t | U = u] = P_C.$$

However the usual frequentist approach is to calculate the unconditional significance level

$$\Pr[T > t] = N(N-1)\, B(\tfrac{2-t}{1-t}, N-1),$$

where $B(\cdot,\cdot)$ is the beta function, Likes (1966).

The application of this approach to translated exponential distributions in the presence of censoring is being extensively treated in another paper.

4. Normal Samples--CPD Tests.

First we present a rather simple illustration. Let $Y_1, \ldots, Y_N$ be $N(\theta,1)$ and assume θ is $N(\beta,\tau^2)$ where β and τ^2 are presumed known. Then the predictive distribution of Z is $N(a,b_0^2)$ where

$$a = \frac{\tau^2\bar{y} + \frac{1}{N}\beta}{\tau^2+N^{-1}} \quad \text{and} \quad b_0^2 = 1 + \tau^2(N\tau^2+1)^{-1}.$$

Hence for Z based on $y_{(i)}$ rather than $y^{(N)}$ we have $Z \sim N(a_i,b^2)$ where

$$a_i = \frac{\tau^2\bar{y}_{(i)} + \frac{1}{N-1}\beta}{\tau^2 + \frac{1}{N-1}}, \quad b^2 = 1 + \tau^2[1 + (N-1)\tau^2]^{-1},$$

where $(N-1)\bar{y}_{(i)} = \sum_{j\neq o} y_j$.

In this instance all the discordancy indices previously discussed select that y_i which maximizes

$$\left(\frac{y_i-a_i}{b}\right)^2 = v_i^2.$$

To implement the unconditional test we note that $X' = (X_1, \ldots, X_N)$ for

$$X_i = Y_i - a_i$$

is a multivariate $N(\mu, \Sigma|\theta)$ where

$$\mu = \frac{(\theta-\beta)e}{(N-1)\tau^2+1}$$

$$e' = (1, \ldots, 1)$$

and Σ is a know matrix with equal diagonal elements and equal off diagonal elements. One then would find the distribution of the $\max|Y_i - a_i|$ or

$\max_i \left(\frac{Y_i - a_i}{b}\right)^2 = V_C^2$ conditional on θ and integrate this over the distribution of θ

to obtain the unconditional distribution of V_C^2. The calculation of

$$\Pr[V_C^2 \geq v_C^2] = P$$

would then result in the significance level. It is clear that even for such a simple normal case, the calculation can be quite arduous especially in cases where the variance is also unknown.

We suggest using the conditional approach to test discordancy here and for the subsequent normal sampling that we shall discuss. Now conditional on $y_{(i)}$

the discordancy index

$$V^2 = \left(\frac{Z - a_i}{b}\right)^2 - X_1^2 .$$

A conditional significance test level can be calculated as

$$\Pr[V^2 \geq v_C^2 | V^2 \geq v_{C-1}^2, y_{(C)}] = P_C$$

or

$$P_C = \frac{1 - F(v_C^2)}{1 - F(v_{C-1}^2)}$$

where $F(\cdot)$ is the distribution function of a X_1^2 random variable.

A particular case of greater interest is when the random sample is from a $N(\mu,\sigma^2)$ and μ and σ^2 are unknown. To make this situation closely correspond to a conventional frequency analysis we let the prior density be

$$g(\mu,\sigma^2) \propto 1/\sigma^2,$$

noting that now the unconditional predictive test is unavailable. Again computation of any of the reasonable discordancy indices lead to the selection of that y_i which maximizes $(y_i-\bar{y}_{(i)})^2$ or

$$\frac{(y_i-\bar{y}_{(i)})^2(N-1)}{Ns^2_{(i)}}$$

which will yield the same y_C since $s^2_{(C)} \le s^2_{(i)}$ where $(N-1)\bar{y}_{(i)} = \sum_{j\neq i} y_j$ and

$(N-2)s^2_{(i)} = \sum_{j\neq i} (y_j-\bar{y}_{(i)})^2.$

We note that the predictive distribution of the "discordancy index"

$$W^2 = \frac{(Z-\bar{y}_{(i)})^2(N-1)}{Ns^2_{(i)}}$$

conditional on $y_{(i)}$ has an $F(1,N-2)$ distribution. Hence to test for discordancy of y_C we can calculate the conditional or adjusted significance level

$$\begin{aligned} P_C &= \Pr[W^2 > w_C^2 | W^2 > w_{C-1}^2, y_{(C)}] \\ &= \frac{1 - F(w_C^2)}{1 - F(w_{C-1}^2)} \end{aligned}$$

where w_{C-1}^2 is the observed second largest standardized deviate and $F(\cdot)$ is the distribution of an F variate with 1 and N-2 degrees of freedom.

For an application of this to a data set that gives results indistinguishable from the rather complex frequency calculation see Geisser (1987). In situations where discordancy indices for the largest y_C and second largest y_{C-1} appear to be appreciably different from the rest, a joint test will depend on the predictive distribution of Z_1 and Z_2, a pair of future observations. This is easily shown to be bivariate student with density function

$$f(z_1,z_2|y_{(i,j)}) \propto \left(1 + \frac{(N-2)t'S^{-1}t}{(N-2)^2 - 1}\right)^{-(N-1)/2}$$

where $y_{(i,j)}$ is the set of observation with y_i and y_j omitted and

$$t' = (z_1-\bar{y}_{(i,j)}, z_2-\bar{y}_{(i,j)})$$

$$(N-2)\bar{y}_{(i,j)} = \sum_{k\neq i,j} y_k$$

$$(N-3)s^2_{(i,j)} = \sum_{k\neq i,j} (y_k-\bar{y}_{(i,j)})^2$$

$$S = s^2_{(i,j)}\begin{pmatrix} 1 & (N-1)^{-1} \\ (N-2)^{-1} & 1 \end{pmatrix}.$$

Further we find that the predictive distribution for

$$W = \frac{(N-2)}{2N-2}T'S^{-1}T \sim F(2,N-3)$$

so that if there were reason to suspect two observations prior to calculating discordancy indices

$$Pr[W \geq w] = P_{ij}$$

where w is the observed value of W for y_i and y_j.

When a discordancy index is calculated here for two observations it is natural to choose the most discordant pair as that y_i and y_j that minimizes P_{ij}. Another would be the one that minimized the bivariate CPO, $d_{ij} = f(y_i, y_j | y_{(i,j)})$. Both will lead to the same most discordant pair.

A possible adjusted test then is

$$P_C = \Pr[W \geq w_C | W \geq w_{C-1}] = \frac{1 - F(w_C, 2, N-3)}{1 - F(w_{C-1}, 2, N-3)}$$

where $F(\cdot, 2, N-3)$ is the distribution function of an $F(2,N-3)$ variate and W_{C-1} is the second largest observed value of W where both arguments involve two observations other than those contained in w_C. Clearly other regions may also be used to make significance calculations but none are likely to be as simple as the proceeding one.

The extension of these CPD tests to regression analysis, Geisser (1987), and to conjugate priors presents no intrinsic difficulty.

5. Acknowledgment

This work was supported by National Science Foundation Grant DMS-8601314 and National Institutes of Health Grant GM 25271.

REFERENCES

Aitchson, J. and Dunsmore, I.R. (1975). Statistical Prediction Analysis, Cambridge: University Press.

Barnard, G.A. (1962). Prepared Contribution. The Foundations of Statistical Inference. L.J. Savage and Other Contributors, edited by G.A. Barnard and D.R. Cox, Methuen & Co., London, pp. 39-49.

Barnett, V. and Lewis, T., (1978). Outliers in Statistical Data. John Wiley and Sons.

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness, J. Roy. Statist. Soc. B, 143, 383-430.

Dixon, W. J. (1950). Analysis of extreme values, Annals of Math. Statistics 21, 488-506.

______ (1951). Ratios involving extreme values, Annals of Math. Statistics 22, 68-78.

Fisher, R.A. (1956). Statistical Methods and Scientific Inference, Oliver and Boyd, Edinburgh.

Geisser, S. (1980). Discussion on Sampling and Bayes' inference in scientific modelling and robustness, by G.E.P. Box. J. Roy. Statist. Soc, A, 143, 416-417.

________ (1985). On the predicting of observables: A selective update, in: Bayesian Statistics 2, with discussion. ed. J.M. Bernardo et al. North-Holland, 203-230.

________ (1987). Influential observations, diagnostics and discordancy tests. Journal of Applied Statistics, 14, 2, 133-142.

Johnson, W. and Geisser, S. (1982). Assessing the predictive influence of observations. Statistics and Probability Essays in Honor of C.R. Rao. (G. Kallianpur, P.R. Krishnaiah, and J. Ghosh, eds.) Amsterdam: North Holland, 353-358.

________ (1983). A predictive view of the detection and characterization of influential observations in regression analysis. J. Amer. Statist. Assoc., 78, 137-144.

Likes, J. (1966). Distribution of Dixon's statistics in the case of an exponential population, Metrika, 11, 46-54.

Smith, A.F.M. and Pettit, L.I. (1985). Outliers and influential observations in linear models, Bayesian Statistics 2, ed. J.M. Bernardo et al. North-Holland, 473-494.